Construction d'un système lexical multilingue, libre de droits, centré sur le français et le japonais via des méthodes automatiques et contributives

Mathieu Mangeot

Construction d'un système lexical multilingue, libre de droits, centré sur le français et le japonais via des méthodes automatiques et contributives

Projet de recherche au Japon

Mathieu MANGEOT-NAGATA
Laboratoire GETALP-LIG
41 rue des mathématiques, BP 53
F-38041 Grenoble CEDEX 9, FRANCE
Mathieu.Mangeot@imag.fr

Introduction

Ce projet de recherche se situe dans le domaine du traitement automatique des langues (TAL), à la croisée de l'informatique et de la linguistique, plus précisément sur la lexicographie et la lexicologie multilingues.

Lors d'un premier long séjour au Japon de novembre 2001 à mars 2004, nous avions fait le constat que les ressources lexicales français-japonais disponibles sur le Web étaient quasi inexistantes. Ce qui avait donné naissance au projet Papillon de construction d'une base lexicale multilingue à structure pivot (Sérasset et al., 2001). Depuis, des progrès ont été faits dans plusieurs domaines (technique, théorique, social) (Mangeot, 2006) mais la production concrète de données a très peu progressé. D'autre part, la réutilisation de ressources lexicales est à la mode (désambiguïsation lexicale, utilisation de ressources en source ouverte (Wiktionary, dbpedia), fusion avec des ontologies, etc.). Même si elles permettent de consolider et d'élargir la couverture des ressources existantes, ces expériences partent toujours de données créées à la main par des lexicographes.

Partant de ce constat, nous avons défini le projet suivant qui consiste à construire un système lexical multilingue riche d'informations avec priorité sur le couple de langues français-japonais. La construction se fera d'une part par la réutilisation de ressources existantes (lexiques franco-japonais, Wiktionary) et leur exploitation automatique (réification de liens de traduction, désambiguïsation de sens de mots) et d'autre part par des contributeurs bénévoles travaillant en communauté sur le Web. Ceux-ci seront amenés à contribuer soit via des jeux lexicaux sérieux, soit directement sur les articles de dictionnaire en fonction de leur niveau d'expertise et de leurs connaissances dans le domaine de la lexicographie ou de la traduction bilingue.

Les ressources ainsi produites seront libres de droits et destinées à être utilisées aussi bien par des humains via des dictionnaires bilingues classiques que par des machines pour des outils de traitement automatique de la langue (analyse, traduction automatique, etc.).

Nous effectuerons d'abord un bref état des lieux de la lexicographie bilingue en général puis du couple français-japonais en particulier. Nous présenterons ensuite les récentes avancées dans le domaine de la construction de ressources lexicales en ligne. Puis, nous décrirons plus en détails le système lexical que nous envisageons de construire. Nous terminerons par une description des étapes nécessaires à cette construction.

1 État des lieux de la lexicographie bilingue

La lexicographie se trouve aujourd'hui à un tournant, entamé à la fin du XX^{ème} siècle avec l'essor de l'informatique. Les dictionnaires papier ont de plus en plus de mal à trouver un public et leurs versions électroniques sont loin d'être des succès commerciaux.

La principale difficulté actuelle de la lexicographie bilingue réside dans les coûts prohibitifs de construction de grandes quantités de données. Par exemple, le projet Electronic Dictionary Research

(EDR, 1993) dont le but était de construire un dictionnaire japonais-anglais a nécessité plus de 1 200 hommes années de travail. Son prix de vente de 84 000 € environ, est très inférieur aux coûts réels de la construction, coûts qui ne seront probablement jamais rentabilisés.

De toutes façons, ces coûts sont trop élevés pour un particulier. De ce fait, seules des institutions peuvent l'acquérir. D'autre part, les données fournies à ce prix ne sont utilisables que par certains systèmes de traduction automatique fondés sur des techniques particulières.

Face à ces coûts difficilement gérables, les maisons d'édition finissent par vivre sur leurs acquis et ne proposent principalement que des nouvelles éditions de dictionnaires existants. Rares sont les éditeurs à avoir le courage de se lancer dans la réalisation d'un nouveau dictionnaire bilingue de qualité en partant de zéro.

D'autre part, même dans les dictionnaires les plus complets, on constate quasiment toujours un manque d'informations en particulier concernant les collocations. Les rares ressources qui en tiennent compte ne le font pas de manière systématique.

Malgré l'arrivée d'Internet, il existe à l'heure actuelle peu de ressources lexicales de bonne qualité disponibles gratuitement en ligne. La plupart sont en fait des lexiques bilingues faits par des bénévoles non spécialistes en lexicographie.

La lexicographie multilingue en tant que telle n'en est en fait qu'à ses débuts. En effet, il n'existe pas vraiment de moyen d'imprimer un vrai "dictionnaire multilingue". Il est par contre tout à fait possible de trouver des bases terminologiques multilingues (comme IATE) ou bien, à la rigueur des petits lexiques ou livres de phrases multilingues.

2 État de l'art des ressources lexicales bilingues japonais

Bien que le français et le japonais soient considérées comme des langues bien dotées au niveau des outils et des ressources linguistiques, le couple français-japonais est considéré comme un couple de langues peu doté. Il existe en effet peu de ressources lexicales bilingues électroniques de qualité et libres de droits. Les corpus bilingues alignés et les systèmes de traduction automatique français-japonais sont logiquement tout aussi rares.

Pour des raisons historiques autant que pratiques, les japonais ont mis rapidement l'accent sur l'anglais. Le couple anglais-japonais est donc l'un des mieux dotés à l'heure actuelle avec des ressources très conséquentes comme le dictionnaire EDR (1993) et des systèmes de traduction automatique parmi les plus performants.

2.1 Dictionnaires éditoriaux français–japonais

Les dictionnaires japonais-français existants de bonne qualité sont des dictionnaires éditoriaux qui n'existent qu'au format papier ou compilé dans des dictionnaires électroniques (denshi-jishou). Il n'existe pas d'interface de consultation en ligne.

2.1.1 Dictionnaires français→japonais

Le Dico (Hakusuisha, 1993) contient 34 000 entrées.

Le Crown (Sanseido) contient 47 000 entrées.

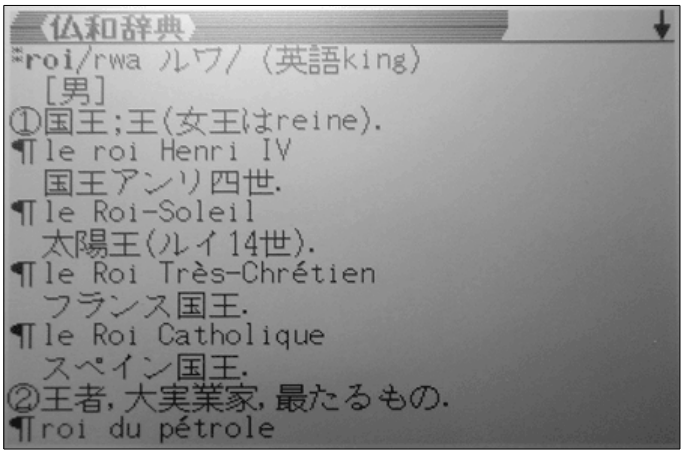

Figure 1 : Capture d'écran du dictionnaire Crown en version électronique

2.1.2 Dictionnaires japonais→français

Le Royal (Obunsha, 1992) contient 42 000 entrées.

Le Concise (Sanseido) contient 38 000 entrées.

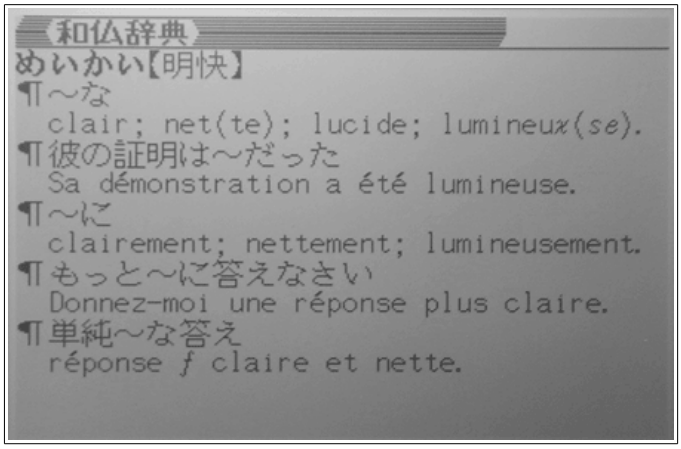

Figure 2 : Capture d'écran du dictionnaire Concise en version électronique

Conclusion

Avantages : ces dictionnaires sont très complets avec des définitions et exemples d'usage.

Inconvénients : payants, ces dictionnaires sont protégés par copyright. Ils ne sont ni modifiables ni réutilisables dans d'autres projets. Les dernières versions éditées datent d'ailleurs d'une vingtaine d'années. De plus, un bon nombre de ces dictionnaires ont été conçus pour des utilisateurs japonophones (Le Dico, Le Royal, etc). Il n'y a donc pas de transcription latine des mots japonais. Leur lecture nécessite donc la maîtrise des kanji, ce qui réduit considérablement leur utilisabilité.

2.2 Projets Wiktionary

Le Wiktionary français compte actuellement 2,2 Mo d'entrées dont 1,2 Mo d'entrées françaises et avec un peu moins de 7 000 traductions en japonais dont une sur deux environ est un nom propre (c'est souvent une simple transcription en syllabaire japonais du mot-vedette). On trouve également quelques traductions reprises du site dictionnaire-japonais.com (voir plus bas). Les traductions sont indiquées au niveau de l'entrée et non des sens de mot. Il n'y a aucune description de contexte de

traduction (glose, exemples), ni d'informations sur la traduction japonaise (classe grammaticale, etc).

Le Wiktionary japonais compte 83 000 entrées dont 26 000 entrées japonaises et 2 800 entrées françaises traduites en japonais (on y trouve des formes fléchies ou formes verbales conjuguées, par exemple 32 entrées pour le verbe « aimer »). La couverture est très insuffisante.

Les projets Wiktionary sont intéressants et à la mode mais ils ont plusieurs limitations :

- La structure des articles est libre. Il n'est pas possible d'utiliser la même microstructure précise pour tous les articles.
- Même s'il est possible de décrire, dans le Wiktionary d'une langue A, un sens de mot d'une langue B dans la langue A, l'interface de départ n'est pas conçue pour rédiger des dictionnaires bilingues. Par exemple, la description du lien inverse langue A→langue B doit être fait à la main dans le projet Wiktionary de la langue B.
- Il n'est pas non plus possible d'ajouter automatiquement des données existantes provenant d'autres sources pour construire un brouillon à raffiner ultérieurement.
- Les contributions sont anonymes. Il n'est pas possible d'utiliser un niveau de qualité des données ou un système de relecture/validation.

2.2 Ressources japonais–autre langue en ligne

2.2.1 Dictionnaire japonais-anglais : Jmdict

Le JMdict[1] (Japanese-Multilingual Dictionary) (Breen, 2004) est un projet mené par Jim Breen. Il contient 165 000 entrées japonais-anglais avec des ajouts de traductions dans d'autres langues : allemand (provenant de WaDokuJiten), 31 000 équivalents français (provenant du dico FJ), russe, etc.

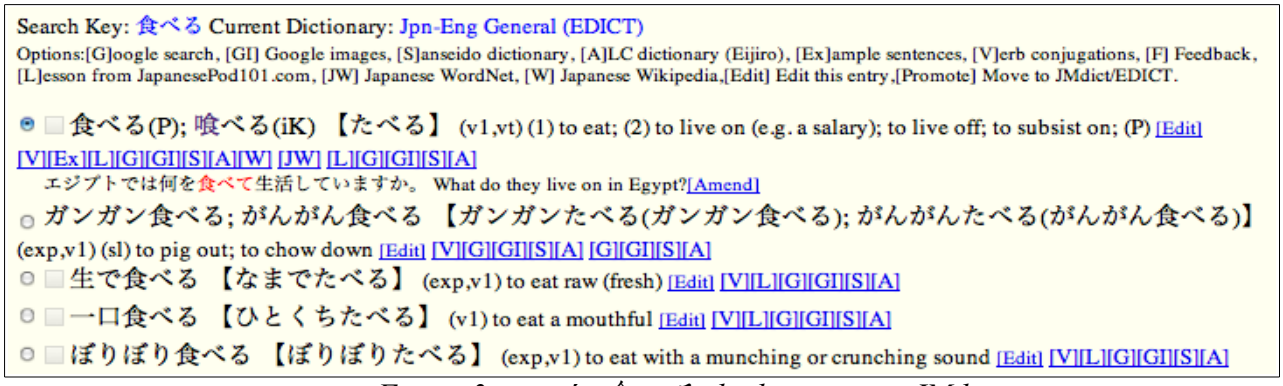
Search Key: 食べる Current Dictionary: Jpn-Eng General (EDICT)
Options:[G]oogle search, [GI] Google images, [S]anseido dictionary, [A]LC dictionary (Eijiro), [Ex]ample sentences, [V]erb conjugations, [F] Feedback, [L]esson from JapanesePod101.com, [JW] Japanese WordNet, [W] Japanese Wikipedia,[Edit] Edit this entry,[Promote] Move to JMdict/EDICT.
食べる(P); 喰べる(iK) 【たべる】 (v1,vt) (1) to eat; (2) to live on (e.g. a salary); to live off; to subsist on; (P) [Edit]
[V][Ex][L][G][GI][S][A][W] [JW] [L][G][GI][S][A]
エジプトでは何を食べて生活していますか。 What do they live on in Egypt?[Amend]
ガンガン食べる; がんがん食べる 【ガンガンたべる(ガンガン食べる); がんがんたべる(がんがん食べる)】
(exp,v1) (sl) to pig out; to chow down [Edit] [V][G][GI][S][A] [G][GI][S][A]
生で食べる 【なまでたべる】 (exp,v1) to eat raw (fresh) [Edit] [V][L][G][GI][S][A]
一口食べる 【ひとくちたべる】 (v1) to eat a mouthful [Edit] [V][L][G][GI][S][A]
ぼりぼり食べる 【ぼりぼりたべる】 (exp,v1) to eat with a munching or crunching sound [Edit] [V][L][G][GI][S][A]

Figure 3 : entrée 食べる du dictionnaire JMdict

Avantages : ressource à large couverture, libre de droits et disponible gratuitement au téléchargement. Elle est aussi régulièrement révisée et complétée.

Inconvénients : dictionnaire unidirectionnel japonais→autre langue. Il n'existe pas de dictionnaire inverse anglais→japonais. La microstructure est limitée : les contextes de traduction ne sont pas décrits. Il manque également une définition et des exemples.

2.2.2 Dictionnaire japonais-allemand : WaDokuJiten

Le WaDokuJiTen[2] de Ulrich Apel (Apel, 2002) est constitué de plus de 280 000 entrées. Sa large couverture ainsi que sa microstructure sont plus développées que le JMdict.

1 http://www.csse.monash.edu.au/~jwb/jmdict.html
2 http://www.wadoku.de

Nr.	Japanisch	Lesung	Deutsch	Worttyp	
1	食べる	たべる	[1] essen; speisen; zu sich nehmen; fressen; probieren. [2] leben von.	下一他	
2	食べるのを遠慮する	たべるのをえんりょする	nicht essen.	サ変自	
3	食べるとしゃきしゃきする	たべるとしゃきしゃきする	beim essen knusprig sein.	サ変自	

Figure 4 : entrée 食べる du dictionnaire WaDokuJiTen

Avantages : plus complet que le JMdict en terme de couverture et d'informations, libre de droits et disponible gratuitement au téléchargement.

Inconvénients : tout comme le JMdict, le dictionnaire est unidirectionnel. Il ne comporte pas non plus d'exemples d'usage pour illustrer les contextes de traduction.

Ce dictionnaire est à ce jour la ressource la plus complète japonais-autre langue disponible gratuitement en téléchargement. Il constitue un objectif à atteindre pour notre ressource en termes de couverture.

2.3 Ressources français-japonais disponibles en ligne

2.3.1 Projet Dico FJ

Le projet dico FJ, précurseur dans le domaine, a été lancé début 2000 par Jean-Marc Desperrier (Desperrier, 2002). Il contient un peu plus de 10 000 entrées provenant de traduction du dictionnaire japonais-anglais JMdict de Jim Breen. Il n'y a pas eu d'évolution depuis 2003.

Avantages : libre de droits et disponible gratuitement au téléchargement.

Inconvénients : en plus des inconvénients du JMdict, on trouve des erreurs de traduction dues au fait que certains contributeurs maîtrisant mal le japonais ont traduit directement les traductions anglaises au lieu des entrées japonaises, ce qui augmente le nombre de contresens.

2.3.2 dictionnaire-japonais.com

Le projet dictionnaire-japonais.com[3] contient un peu plus de 28 000 mots. Il constitue un net progrès par rapport aux autres projets de dictionnaire japonais-français en ligne. Chaque utilisateur peut contribuer directement en rajoutant des entrées. La communauté de contributeurs semble assez active comme en témoigne l'activité sur le forum du projet. Les informations disponibles pour chaque entrée sont relativement limitées à un "type grammatical", une "catégorie" (domaine), un registre de langue, et parfois une "origine du mot" (étymologie).

3 http://www.dictionnaire-japonais.com

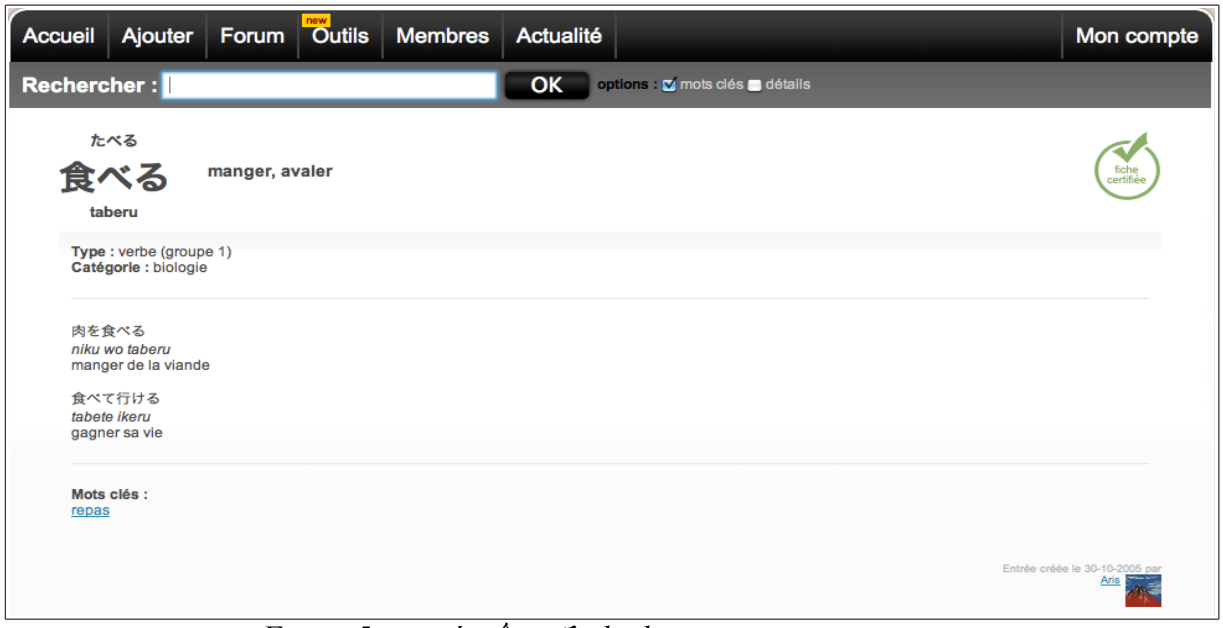

Figure 5 : entrée 食べる du dictionnaire-japonais.com

Avantages : disponible en ligne, couverture un peu plus large que le dico FJ, communauté de active de contributeurs bénévoles.

Inconvénients : en plus des inconvénients du dico FJ, il semble que les fichiers sources des données ne soient pas disponibles au téléchargement.

2.4 Lexiques issus de systèmes de traduction

Le projet UNL[4] définit un langage pivot sémantique, le langage UNL. Pour chaque langue, il est ensuite possible de programmer des enconvertisseurs langue→UNL et des déconvertisseurs UNL→langue. Les partenaires français et japonais du projet ont chacun produit un lexique pour leurs systèmes de conversion. Ces lexiques sont consultables en ligne[5].

Le lexique japonais-UNL comporte 1 266 694 entrées. Voici l'entrée japonaise 食べる(manger) :

[食べる]{} "eat(icl>food)";

Le lexique français-UNL comporte 520 305 entrées. Voici les entrées correspondant au verbe « manger » :

[manger]{AUX(AVOIR),CAT(CATV),GP1(DE),VAL1(GN)}"consume(icl>event)";

[manger]{AUX(AVOIR),CAT(CATV),VAL1(GN)}"eat";

Avantages : ces lexiques ont une couverture assez large et peuvent être réutilisés pour construire un squelette de dictionnaire.

Inconvénients : la microstructure est très réduite : un lemme en français ou japonais et son équivalent en UNL avec, parfois des informations provenant de systèmes de traduction comme la classe grammaticale (CAT). Ces lexiques ne sont pas disponibles en téléchargement libre.

2.5 Conclusion

Les lexiques japonais-français disponibles en ligne sont d'une part de petite taille et d'autre part tous orientés japonais vers français.

La plupart des dictionnaires français-japonais manque aussi d'informations spécifiques au japonais. Il n'existe par exemple pas à notre connaissance de dictionnaire présentant à la fois les kanji

4 http://www.undl.org/
5 http://www.undl.org/uwgate/

(idéogrammes), les kana (syllabaires) et le romaji (transcription en alphabet romain). Les dictionnaires sans romaji sont destinés aux japonophones. Les dictionnaires sans kanji sont destinés aux francophones débutants. Il manque aussi certaines informations importantes telles que les compteurs (on ne compte pas de la même manière les objets : une voiture = ichi dai, un chien = ippiki, etc) ou les niveaux de langue.

En conclusion, pour un usage personnel, il est possible de trouver des dictionnaires imprimés (ou leur version électronique) d'assez bonne qualité à condition de savoir lire les kanji; mais lorsque l'on cherche un dictionnaire gratuit ou une ressource réutilisable dans d'autres outils, il n'y a bien souvent pas d'autre choix que d'utiliser un dictionnaire anglais-japonais, ce qui, on le sait, ne peut que multiplier les erreurs de compréhension et de traduction.

3 Avancées dans la construction de ressources en ligne

Notre thèse (Mangeot, 2001) définit un certain nombre de fondements théoriques dans le domaine. Les premiers pas concrets ont été réalisés avec le projet Papillon : définition des besoins, des structures et types d'information souhaités, mise en œuvre de la plate-forme Jibiki de manipulation de ressources lexicales en ligne. Ont suivi une réflexion sur les moyens de contribuer en ligne de manière indirecte via des jeux sérieux à travers le projet JeuxDeMots. Plusieurs projets de construction de ressources ont également été menés à l'aide de la plate-forme Jibiki mais aucun n'était basé sur le couple franco-japonais. Le projet GDEF (Mangeot et al., 2006) toujours en cours a pour objectif de construire un dictionnaire éditorial estonien-français de haute qualité pour traducteurs. Le projet LexAlp (Sérasset et al, 2006) a produit une terminologie multilingue (allemand, français, italien, slovène) à structure pivot pour le vocabulaire de la convention alpine. Le projet MotÀMot (Mangeot et al. 2010) s'est concentré sur la récupération d'un dictionnaire existant français-khmer.

3.1 Sur les aspects contributifs : les projets Wikipedia et Wiktionary

L'encyclopédie en ligne contributive Wikipedia a rencontré un gros succès incontestable. Même si les projets Wiktionary se développent, le succès n'est pas encore au rendez-vous pour certaines langues (26 000 pour le japonais) et ils sont très peu utilisés pour construire des dictionnaires bilingues. Le projet Wiktionaryz, qui prétendait parer aux défauts de Wiktionary n'a pas non plus eu l'effet escompté.

Une hypothèse pour expliquer ce problème est celui de la motivation. En effet, lorsqu'une personne contribue à un article de Wikipedia, elle est récompensée par la renommée. Elle sera ensuite reconnue comme un expert dans son domaine. Cela n'est pas possible avec un dictionnaire. Les contributions portent sur des petites parties d'informations très ciblées et sont de ce fait anonymes. D'autre part, il y a un aspect technique lié à la structure. Un article d'encyclopédie a une structure plus ou moins libre tandis qu'une entrée de dictionnaire doit suivre une structure très précise (mot-vedette, informations grammaticales, blocs sémantiques, bloc de traduction, blocs d'exemples, etc.). Il n'est donc pas possible de réutiliser une plate-forme wiki pour construire un dictionnaire avec une structure bien définie.

Une fois acceptée l'idée que rédiger des entrées de dictionnaire n'est pas aussi plaisant que travailler sur un article de Wikipédia, il faut trouver des solutions pour motiver une communauté de bénévoles à contribuer à un dictionnaire. Les jeux sérieux lexicaux constituent une première piste. Il faut aussi mettre en valeur les contributeurs à travers par exemple un tableau des meilleurs contributeurs du mois. Et enfin, l'exploitation de réseaux communautaires tels que Facebook devraient aussi apporter de l'eau au moulin.

3.2 Sur les aspects techniques : la plate-forme Jibiki

Jibiki (Mangeot et al., 2004) est une plate-forme générique en ligne pour manipuler des ressources

lexicales avec gestion d'utilisateurs et groupes, consultation de ressources hétérogènes et édition générique d'articles de dictionnaires. C'est un site Web communautaire développé au départ pour le projet Papillon. La plate-forme est programmée entièrement en Java, basée sur l'environnement "Enhydra". Toutes les données sont stockées au format XML dans une base de données (Postgres). Ce site Web propose principalement deux services : une interface unifiée permettant d'accéder simultanément à de nombreuses ressources hétérogènes (monolingues, dictionnaires bilingues. bases multilingues, etc.) et une interface d'édition spécifique pour contribuer directement aux dictionnaires disponibles sur la plate-forme.

L'éditeur est basé sur un modèle d'interface HTML instancié avec l'article que l'on veut éditer. Le modèle peut être généré automatiquement depuis une description de la structure de l'entrée à l'aide d'un schéma XML. Il peut être modifié ensuite pour améliorer le rendu à l'écran. La seule information nécessaire à l'édition d'un article de dictionnaire est donc le schéma XML représentant la structure de cette entrée. De plus, il est possible d'éditer n'importe quel type de dictionnaire s'il est encodé en XML.

Plusieurs projets de construction de ressources lexicales ont utilisé ou utilisent toujours cette plate-forme avec succès. C'est le cas par exemple du projet GDEF de dictionnaire bilingue estonien-français[6]. Le code de cette plate-forme est disponible gratuitement en source ouverte en téléchargement depuis la forge du laboratoire LIG[7].

3.3 Sur la collecte de données via des jeux sérieux : les projets JeuxDeMots

JeuxDeMots[8] (Lafourcade & Joubert, 2008) est une tentative de réponse au problème de contribution anonyme. Ce projet a pour but de construire un réseau lexical riche et évolutif, qui peut être comparé à un certain degré à la fameuse base WordNet (Miller et al., 1990). Le principe est le suivant : une partie nécessite 2 joueurs. Lorsqu'un joueur A débute une partie, une consigne concernant un type de compétence (synonymes, contraires, domaines, etc.) est affichée, ainsi qu'un mot M tiré aléatoirement dans une base de mots. Le joueur A a alors un temps limité pour répondre en donnant des propositions répondant, selon lui, à la consigne appliquée au mot M. Ce même mot, avec cette même consigne, est proposé à un autre joueur B ; le processus est identique.

Les deux demi-parties, celle du joueur A et celle du joueur B, ne sont pas simultanées, mais asynchrones. Pour toute réponse commune dans les propositions de A et B, ces deux joueurs gagnent un certain nombre de points. La structure du réseau lexical que nous cherchons ainsi à obtenir s'appuie sur les notions de nœuds et de relations entre nœuds pour construire un réseau du type de (Polguère, 2006). Chaque nœud du réseau est constitué d'une unité lexicale (terme ou expression) regroupant toutes ses lexies et les relations entre nœuds traduisent des fonctions lexicales. La première version du jeu pour le français a été lancée en juillet 2007. Il existe aussi des versions anglaises, arabes, japonaises, vietnamiennes, thaï et plus récemment portugaise (Mangeot et al., 2012). Elles sont disponibles sur le Web.

4 Description du système lexical à construire

4.1 Microstructure des articles basée sur la théorie sens-texte

La microstructure des articles composant les volumes monolingues est une simplification de celle du projet Papillon. Chaque article est cette fois basé sur le vocable. Un vocable étant soit un regroupement de lexies (sens de mot), soit une locution.

Les lexies sont constituées d'un nom, des propriétés grammaticales, d'une formule sémantique qui peut être vue comme une définition formelle - dans le cas d'une lexie, prédicative, la formule décrit

6 http://www.estfra.ee/
7 http://jibiki.ligforge-imag.fr/
8 http://jeuxdemots.org/

le prédicat et ses arguments et on trouve aussi le régime qui décrit la réalisation syntaxique des arguments - , puis d'une liste de fonctions lexico-sémantiques - il y a 56 fonctions de base applicable à toute langue et pouvant se combiner entre elles -, d'une liste d'exemples et enfin d'une liste d'expressions idiomatiques.

Pour faire face aux niveaux de compétences différents selon les contributeurs, l'interface d'édition pourra s'adapter et afficher une granularité d'information adaptée. Par exemple, un contributeur débutant sera invité à renseigner une simple glose pour caractériser une lexie, alors qu'un linguiste expert devra décrire une formule sémantique complète. De même, certains contributeurs seulement auront accès à la liste des fonctions lexicales à remplir.

4.2 Macrostructure pivot via des interfaces bilingues

La macrostructure est également tirée du projet Papillon avec un volume monolingue pour chaque langue et un volume pivot au centre. Cette macrostructure a été expérimentée et validée dans le projet LexAlp[9] (Sérasset et al., 2006) de construction d'une terminologie multilingue pour le vocabulaire de la convention alpine. Ce projet utilise également la plate-forme Jibiki (Mangeot et al., 2004) comme pour son développement et sa consultation en ligne.

Lorsqu'un nouvel article dans une langue A est ajouté, il doit être relié au volume interlingue. Ces liens sont créés soit en réutilisant des dictionnaires bilingues existants langue A→langue B, soit en les ajoutant manuellement à partir d'une traduction. Le lien langue A→langue B devient langue A→pivot→langue B. Si l'article langue B est déjà relié à un autre article langue C, alors l'article langue A bénéficiera lui aussi de ces liens.

Cependant, afin de ne pas dérouter les utilisateurs, ceux-ci contribueront via une interface présentant une vue classique de dictionnaire bilingue. Chaque lien bilingue langue A→langue B ajouté via cette interface sera en fait traduit en arrière plan par la création de deux liens interlingues ainsi que d'une axie représentant le lien de traduction d'origine pour obtenir finalement : langue A→axie pivot→langue B. Cette idée a été utilisée pour le projet MotÀMot[10] (Mangeot et al., 2010) qui a abouti à la construction d'un dictionnaire bilingue français-khmer.

4.3 Établissement des liens bilingues et interlingues

Lorsqu'un contributeur veut ajouter un lien de traduction entre un vocable Va de langue A et un vocable Vb de langue B, il peut établir ce lien à différents niveaux.

La solution idéale est de relier un sens de mot Sa du vocable Va à un autre sens de mot Sb du vocable Vb. Dans ce cas, le lien est bijectif et Sb est donc aussi relié à Sa.

Si le vocable Vb n'a pas encore de sens de mots précis ou si le contributeur n'arrive pas à choisir de sens de mot, il peut relier Sa directement au vocable Vb. Dans ce cas, un nouveau sens de mot Sb' est créé avec un niveau de qualité brouillon et le lien ainsi que les sens de mots sont marqués comme étant à raffiner.

Dans le cas de la récupération de données existantes, il est bien souvent impossible de rattacher une information à un sens de mot précis. Dans ce cas, on ajoute à la fin du vocable Va l'information selon laquelle un des sens de mots de Va peut être relié à un sens de mot de Vb, mais cette information ne sera pas ajoutée à Vb. Elle sera bien sûr marquée comme étant à raffiner d'urgence !

Grâce à la macrostructure pivot, si deux liens langue A→langue B et langue B→langue C existent, alors il sera automatiquement créé un lien langue A→langue C dont le niveau sera de qualité brouillon et marqué comme à réviser.

9 http://217.199.4.152:8080/termbank/LexALP.po
10 http://jibiki.univ-savoie.fr/motamot/

4.4 Niveaux de qualité des données et des contributeurs

Chaque partie d'information de chaque article se verra attribuer un niveau de qualité . Les niveaux s'échelonnent de 1 étoile pour un brouillon (données récupérées dont la qualité n'est pas connue) à 5 étoiles, qualité certifiée par un expert (par exemple, un lien de traduction validé par un traducteur assermenté).

De la même manière, les contributeurs se verront assigner un niveau de compétence (1 à 5 étoiles également). 1 étoile étant le niveau d'un débutant inconnu dans la communauté et 5 étoiles étant le niveau d'un expert reconnu.

Ensuite, lorsqu'un contributeur de niveau 3 révise un article de niveau 2, l'article monte automatiquement au niveau 3. De même, si le travail d'un contributeur est systématiquement validé sans corrections par d'autres contributeurs de niveau supérieur, celui-ci peut passer automatiquement au niveau supérieur au bout d'un certain seuil (par exemple 10 contributions).

Pour aller plus loin, nous envisageons d'analyser le travail des contributeurs. Si une personne contribue massivement par exemple sur un domaine particulier, le système pourra de manière automatique lui envoyer régulièrement des propositions de contribution dans son domaine.

5 Plan de travail prévu

5.1 Collecte de ressources existantes

La première étape du projet est de collecter des ressources existantes que nous sommes autorisés à réutiliser (libres de droits principalement). Il s'agit de ressources lexicales (dictionnaires, lexiques) et corporales (corpus bilingues alignés ou comparables). Ces dernières serviront pour l'extraction d'exemples bilingues.

5.1.1 Dictionnaires disponibles sur Internet

- Le dictionnaire JMdict avec ses 31 000 équivalents français ;
- Les traductions des Wiktionary français et japonais.

5.1.2 Ressources de projets dans le domaine du TAL

- Données du projet Sakura-survitra en cours avec l'université de Kyoto ;
- Lexiques français-UNL et japonais-UNL du projet UNL avec l'Université des Nations Unies et Tokyo Soft ;
- Données pré-terminologiques issues de la thèse de Mohammad Daoud co-encadré avec Pr. Kyo Kageura, Université de Tokyo (Daoud et al. 2010), (Daoud, 2010) ;
- Données éventuelles d'autres partenaires japonais (Université Hosei, Institut National d'Informatique, Laboratoire du Pr. Fuji Ren à l'université de Tokushima, Laboratoire du Pr. Yoshinori Sagisaka à l'université Waseda, Laboratoire du Pr. Yves Lepage à l'université de Waseda, etc.)

5.1.3 Lexiques personnels des traducteurs-interprètes

En l'absence de dictionnaire de référence à large couverture ou de bases terminologiques spécialisées, il est fréquent que les traducteurs-interprètes construisent leurs propres lexiques spécialisés le plus souvent à l'aide de fichiers Excel. Ces ressources pourraient être réutilisées dans le cadre du projet.

5.2 Production d'un squelette de dictionnaire à réviser

Une fois collectées, les ressources doivent être converties dans un format unique puis fusionnées automatiquement. Deux entrées ayant le même mot-vedette et la même classe grammaticale sont fusionnés en un seul article avec plusieurs sens de mot. Les liens de traduction doivent être réifiés : à partir d'un article français avec une traduction japonaise Fj, un article japonais J et une entrée pivot reliant les deux articles A sont générés (Fj ⇒ F→ A←J).

La deuxième étape consiste à fusionner éventuellement les différents sens de mot d'une entrée résultant d'une précédente fusion. Il s'agit d'utiliser ici des techniques de désambiguïsation lexicale (Navigli, 2009) comme les algorithmes à colonies de fourmis (Schwab et al. 2011).

Lorsqu'une première version de la ressource est constituée, celle-ci est ensuite mise en ligne pour être corrigée et enrichie d'une part semi-automatiquement par des programmes et d'autre part manuellement par une communauté de contributeurs bénévoles.

5.3 Enrichissement semi-automatique de la ressource

Nous prévoyons de baser l'enrichissement de la ressource par des données issues des JeuxDeMots via deux méthodes principales :

1. Par élicitation lexicale : des récentes recherches ont prouvé que les «jeux sérieux lexicaux», tels que JeuxDeMots, sont très utiles pour recueillir des données lexicales monolingues. Des recherches prometteuses sont actuellement menées sur les moyens de développer des jeux bilingues ou multilingues basés sur JeuxDeMots. Alors que JeuxDeMots a déjà produit un grand réseau lexical pour le français, un objectif important de ce projet est d'élargir considérablement le public du JeuxDeMots (existant) japonais[11].
2. Par découverte et traduction de collocation : une part importante des ressources nécessaires est faite de collocations, et en particulier des termes constitués de de plusieurs mots (expressions polylexicales). Des traitement statistiques ont été utilisés avec succès pour trouver ces termes dans les corpus anglais spécialisé (par exemple, dans le corpus Genia), et de proposer des traductions dans d'autres langues.

5.4 Animation d'une communauté de contributeurs autour du projet

Une première conclusion de nos expériences sur les JeuxDeMots est que les projets basés sur le travail de contributeurs bénévoles doivent impérativement être constamment animés sous peine d'être rapidement abandonnés passé les premiers temps de découverte. Une de nos tâches essentielles sera donc d'animer une communauté de contributeurs via la publicité du projet, la modération de forums, les réseaux sociaux, etc. Il sera également nécessaire de recruter des experts pouvant superviser des contributeurs et animer également la communauté. Des outils peuvent également être mis en place pour récompenser des contributeurs (meilleur contributeur du mois, etc.).

Nous prévoyons enfin de reprendre la tâche de communication scientifique autour du projet en organisant un nouveau séminaire Papillon au Japon, en donnant des exposés dans les laboratoires partenaires et en participant à des conférences du domaine.

Références

Apel U. (2002) *WaDokuJT - A Japanese-German Dictionary Database.* Papillon 2002 Seminar, 16-18 July 2002, NII, Tokyo, Japan, 13 p.

Berment V. (2004) *Méthodes pour informatiser des langues et des groupes de langues "peu dotées".* Thèse de nouveau doctorat, Université Joseph Fourier Grenoble I, Grenoble, France, 277 p.

11 http://jeuxdemots.liglab.fr/jpn/

Breen JW. (2004) *JMDict: a Japanese-multilingual dictionary.* In: Coling 2004 workshop on multilingual linguistic resources, Geneva, Switzerland, pp. 71-78.

Daoud M., Kageura K., Boitet C., Kitamoto A. & Mangeot M. (2010) *Multilingual Lexical Network from the Archives of the Digital Silk Road.* Proc. of OntoLex 2010, 6thWorkshop on Ontologies and Lexical Resources, hosted by COLING 2010, Beijing, 22 August 2010, 9 p.

Daoud M. (2010) *Utilisation de ressources non conventionnelles et de méthodes contributives pour combler le fossé terminologique entre les langues en développant des "préterminologies" multilingues.* Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, 192 p.

Desperrier J-M. (2002) Analyze [sic] of the results of a collaborative project for the creation of a Japanese-French dictionary. In: Proceedings of Papillon 2002 Seminar, Tokyo, Japan.

EDR (1993) *EDR Electronic Dictionary Technical Guide.* Project Report, n°-042, Japan Electronic Dictionary Research Institute Ltd., 16 August 1993, 144 p.

Lafourcade M. & Joubert A. (2008) *JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes.* In JADT 2008 : 9es Journées internationales d'Analyse statistique des Données Textuelles, Lyon, France, pp. 657–666.

Mangeot M. (2001) *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue.* Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, 280 p.

Mangeot M. (2006) *Papillon project : Retrospective and perspectives.* In P. Zweigenbaum, Ed., Acquiring and Representing Multilingual, Specialized Lexicons : the Case of Biomedicine, LREC workshop, Genoa, Italy, 6 p.

Mangeot M., Sérasset G. & Lafourcade M. (2004) *Construction collaborative d'une base lexicale multilingue.* Traitement Automatique des Langues, vol. 44(2), pp. 151–176.

Mangeot M. & Chalvin A. (2006) *Dictionary building with the Jibiki platform : the GDEF case.* In LREC 2006, Genova, Italy, pp. 1666–1669.

Mangeot M. & Sereysethy Touch S. (2010) *MotÀMot project: building a multilingual lexical system via bilingual dictionaries,* SLTU 2010: Second International Workshop on Spoken Languages Technologies for Under-Resourced Languages, Penang, Malaysia, 3-5 May 2010, 6p.

Mangeot M. & Ramisch C. (2012) *A Serious Lexical Game for Building a Portuguese Lexical-Semantic Network.* Proceedings of the ACL 2012 3rd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources and their Applications to NLP, Jeju, Republic of Korea, jul 2012.

Mel'čuk I., Clas A. & Polguère A. (1995) *Introduction à la lexicologie explicative et combinatoire.* Universités francophones et champs linguistiques. Louvain-la Neuve : AUPELF-UREF et Duculot, 256 p.

Miller G. A. et al. (1990) *Introduction to WordNet : an on-line lexical database.* International Journal of Lexicography, 3(4), pp. 235–244.

Navigli R. (2009) *Word Sense Disambiguation: a Survey.* ACM Computing Surveys, 41(2), ACM Press, 2009, pp. 1-69.

Polguère A. (2000) *Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French*. In Proceeding of EURALEX'2000, Stuttgart, pp. 517–527.

Polguère A. (2006) *Structural properties of lexical systems : Monolingual and multilingual perspectives.* In Workshop on Multilingual Language Resources and Interoperability (COLING/ACL 2006), Sydney, Australia, pp. 50–59.

Schwab D., Goulian J. et Guillaume, N. (2011) *Désambiguïsation lexicale par propagation de mesures sémantiques locales par algorithmes à colonies de fourmis.* 18ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011), 27 juin - 1er juillet 2011, Montpellier.

Sérasset G., Brunet-Manquat F. & Chiocchetti E. (2006) *Multilingual Legal Terminology on the Jibiki Platform: The LexALP Project.* COLING/ACL2006 Conference, 17-21 july 2006, Sydney, Autralia.

Sérasset G. & Mathieu Mangeot M. (2001) *Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links.* Proc. of NLPRS 2001, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan, 27-30 November 2001, pp. 119-125.